Tufts University

EM-0212: Applied Data Science
03-10-2019

Data Project Deliverable 3: Datasheet for Harpswell Property Tax Data

Shea T. Nelson
shea.nelson@tufts.edu

Motivation for Data Creation

Why was the dataset created?

The "2018 Property Tax Commitment" dataset was created in order to give the tax assessor's office enough information about land valuation across the entire town to evaluate the equity and fiscal sustainability of property taxes. Additionally, since it is collected each year, the dataset is also intended to reveal demographic trends for governmental use. Lastly, it simply serves as a record for the town and a reference for any resident or homeowner who desires information on their home's financial history.

The "List of Property Sales" dataset was created to fulfill requirements for public information. In Maine, public records detailing home sales are to be available to the general public and the dataset accomplishes that purpose in an easily accessible and concise manner.

What other tasks could the dataset be used for?

The "2018 Property Tax Commitment" could be used for demographic analysis of the current state of the town. The interrelationships between residency of owner, real estate value, and home location could be investigated, amongst others. Additionally, the records are available over several years so the trends of these areas of interest could be examined.

The "List of Property Sales" has more limited uses. However, since it is also maintained and updated each year, it could be used to reveal trends in home ownership. These could include locations of buyers and sellers, previous residence, sale price vs. home value assessment, effect of tax on value, amongst many others. However, to accomplish this would require in inclusion of the "2018 Property Tax Commitment" dataset. This combination, in fact, forms the basis of this report.

Both datasets contain enough information to be utilized for investigation of several different avenues of town demographics and the trends thereof.

Has the dataset been used for any tasks already?

Every year, the town tax assessor evaluates the most current property tax commitment for the town and attempts to balance, or "equalize," taxes across all properties. In the words of the FAQ associated with the dataset, the purpose of equalization is, "to more equitably distribute the property tax burden based upon valuation."[1]

The "List of Property Sales" datasets however, have no official use noted on the town website beyond simply serving as public information. Additionally, no other apparent uses appear to exist from searching of internet databases.

To the knowledge of the investigator at the time of publication, no dataset under study in this paper has been used for any other tasks beyond those explicitly stated by the Town of Harpswell as outlined in this section.

Tax Assessment FAQ

[1] USA. Harpswell Town Government. Office of the Tax Collector. *Frequently Asked Questions*. October 12, 2018. Accessed March 10, 2019. http://www.harpswell.maine.gov/index.asp?SEC=B8009038-83C1-4082-8297-FFB27C3B4AF0&DE=265AF279-E9ED-4EEA-AFF5-DCC999D755BB&Type=B_BASIC.

Who funded the creation of the dataset?

The "2018 Property Tax Commitment" and "List of Property Sales" datasets were both funded by the Harpswell Town Government through the Office of the Tax Assessor.

Additional Comments

Two datasets are under consideration, "2018 Property Tax Commitment" and "List of Property Sales." The first lists home values and associated tax information while the second lists simply real estate transactions in Harpswell, ME.

Dataset Composition

What are the instances?

For "2018 Property Tax Commitment," each instance consists of general owner identification followed by tabulations of real estate value, exemption values, assessed value, current tax burden, resident status, location, lot location, and lot size. Overall, it essentially contains general information about each piece of real estate and its respective owner in the town.

"List of Property Sales" is comparable to the tax dataset and has instances of original owner identification, new owner identification, date of sale, price, and physical location of real estate.

Both datasets have instances arranged in rows with each row representing an individual instance.

Are there relationships between instances made explicit in the data?

There are no known or apparent relationships between instances made explicit by the data. However, some individuals own or have purchased multiple properties in the town and thus appear in more than one place on one of the datasets.

How many instances of each type are there?

The dataset "2018 Property Tax Commitment," contains 5041 instances containing all information outlined in the section "What are the instances."

The dataset "List of Property Sales," contains 271 instances which contain all information outlined in the section "What are the instances."

An important note to both datasets is that the number of instances changes from year to year due to the construction and demolition of homes, increase or decrease in rate of sale, and parceling of property.

What data does each instance consist of?

For both datasets, each instance consists of text or numerical inputs separated into columns. Each column is a separate element of the information outlined in the section "What are the instances." The data is formatted in Excel and thus capable of being exported in CSV format.

Is everything included or does the data rely on external resources?

Everything is included in both datasets. They are standalone and do not require any access to external resources.

Are there recommended data splits or evaluation measures?

There are no recommended data splits or evaluation measures by the provider of the two datasets under study. An important consideration, however, is that the datasets represent effectively the entirety of the town real estate. Indeed,

since the dataset almost exactly represents n = population, splitting may have undesirable consequences. Furthermore, if the data is to be utilized to investigate demographic context rather than predictive applications, it would be damaging to conduct splitting.

In the case of evaluation measures, since the dataset is a full population rather than a sample, the statistical methods should be associated accordingly.

What experiments were initially run on this dataset?

The data was released without any public experimental results. It is relatively certain that the tax assessor utilized the "2018 Property Tax Commitment" for the equity evaluations outlined in the section, "Has the dataset been used for any tasks already?". However, no publicly available experimental data has been conducted on either dataset.

Additional Comments

The knowledge of the investigator in terms of complex data science is limited and thus this section should be taken with a grain of salt. Although the information contained is accurate, it may not reach the level of sophistication required of further analysis.

Data Collection Process

How was the data collected?

The data for both datasets was collected by the Tax Assessor's Office conducting a review of "deeds, surveys, subdivision plans, taxpayer list declaration forms, building permits and etc."[2] Furthermore, various direct site inspections were conducted and evaluated by tax experts. The data was then entered into a corresponding property database for each dataset. "2018 Property Tax Commitment" was entered into a tax database while "List of Property Sales" was entered into a property change of ownership database. Both datasets were then committed to the Tax Collector for review in 2018.

Who was involved in the data collection process?

The data collection process was mostly limited to the employees and contractors of the Office of the Tax Collector and the Tax Collector herself, Jill Caldwell. However, nearly every homeowner was involved in some sense by the annual reporting and payment of taxes. This information was evaluated by the assessor and compiled into the datasets, largely without change besides the parsing of data strings.

Over what time-frame was the data collected?

The "2018 Property Tax Commitment" under study in this particular paper was collected over the course of the last half of 2017 and the first half of 2018.

"List of Property Sales" is an ongoing and continually updated dataset. The current date range is listed as 04/01/18-03/31/19.

An important aside is that both datasets are compiled every year so in a broader sense they have been collected in electronic form for over a decade.

[2] USA. Harpswell Town Government. Office of the Tax Collector. *Tax Assessor's Office.* Accessed March 10, 2019. http://www.harpswell.maine.gov/index.asp?Type=B_BASIC&SEC={DFEC18E3-AB5A-4A61-AAD7-16B0E368466B}.

How was the data associated with each instance acquired?

It is not known how the data associated with each instance was acquired for either dataset besides the general sources listed in the section "How was the data collected?".

Does the dataset contain all possible instances?

Yes, both datasets contain all possible instances up to the day of analysis.

If the dataset is a sample, then what is the population?

N/A: Both datasets contain the full population and are not samples.

Is there information missing from the dataset and why?

There is no information missing from either dataset to the knowledge of the investigator at the time of publication. The investigator theorizes this to be the result of the legal requirements for sale and ownership of real estate; in essence, if information was missing from the records then the legality of ownership would be challenged until the information was provided.

Are there any known errors, sources of noise, or redundancies in the data?

There are no *known* errors or sources of noise or redundancies in either dataset. However, since human beings are involved in the data provenance there are likely small errors in the reporting of some values. However, this is assumed to be inconsequential in the broader context of each dataset.

Additional Comments

Contacting the Harpswell Town Tax Collector may be helpful in determining more about data collection methods. The lack of information on the matter is more likely due to there not being much demand for it, rather than any sort of malfeasant action.

Assessor's Office Contact Information[3]

Data Preprocessing

What preprocessing/cleaning was done?

Although no direct information has been released about preprocessing for either dataset, both were subject to the interpretation of the tax office employee when determining real estate values. This itself is preprocessing in some sense but no official methods have been released.

Only minimal cleaning was done of the data in both datasets. This results in various idiosyncrasies, (e.g. road having forms rd., ROAD, Rd, rd., etc.) but an essentially true preservation of the original data, simply in organized and consolidated form.

Was the "raw" data saved in addition to the preprocessed/cleaned data?

In the case of both datasets under study, this is a somewhat difficult question to address. The datasets are both amalgamations of multiple data sources (See section "How was the data collected?"). Therefore, the "raw" data is preserved in original form but across a number of different sources.

Additionally, in a broad sense, the datasets themselves represent a sort of

[3] USA. Harpswell Town Government. Office of the Tax Collector. *Contact Information*. Accessed March 10, 2019. http://www.harpswell.maine.gov/index.asp?SEC=0C1A3A74-0691-4004-B4CC-402D17CB4881&DE=F6B3569F-079F-4AAB-81BE-1C54DE46AD2F.

"raw" form for future study and are treated by the town as "raw" data for all intents and purposes.

Is the preprocessing software available?

In some sense, yes. Although some of the preprocessing was done by a human based on their interpretation of home values and other base data, most of the data for both datasets was simply entered into Excel. Excel is readily available to most in a professional setting.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

Yes, the dataset "2018 Property Tax Commitment" sufficiently aggregates real estate values for the tax collector to effectively distribute tax burdens across the town and evaluate tax equity. "List of Property Sales" is easily and publicly accessible in the form of an Excel document online and readably organized, thus fulfilling the requirement of it being available to the public.

Additional Comments

Here again, contacting the Harpswell Town Tax Collector may be helpful in determining more about data preprocessing methods.

Assessor's Office Contact Information[4]

Dataset Distribution

How is the dataset distributed?

Both datasets are distributed on the official Harpswell town website under the Assessor's Office section. The datasets do not have DOI's and are not archived redundantly, at least electronically. However, paper copies are stored at the town office and also publicly available.

Harpswell Assessor's Office Website[5]

When will the dataset be released/first distributed?

The dataset "2018 Property Tax Commitment" has been released yearly for several decades but is first available for each year on the day of assessment: April 1st.

The dataset "List of Property Sales" has been released yearly for several years and is first available for each year starting on the day of assessment: April 1st. However, the dataset is updated every few weeks for one year and is immediately available in updated form when changes are made.

What license (if any) is it distributed under?

Both datasets are required public information and are thus released without any licensing.

Are there any fees or access/export restrictions?

There are no fees or access/export restrictions for either dataset.

Additional Comments

None

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

[4] Ibid.

[5] USA. Harpswell Town Government. Office of the Tax Collector. *Welcome to the Harpswell Assessor's Office*. February 14, 2019. Accessed March 10, 2019. http://www.harpswell.maine.gov/index.asp?SEC=B8009038-83C1-4082-8297-FFB27C3B4AF0&Type=B_BASIC.

The dataset is supported by the Harpswell Town Government and under the supervision of Jill Caldwell. Contact info is available on the Harpswell town website.

Assessor's Office Contact Information[6]

Will the dataset be updated?

The dataset "2018 Property Tax Commitment" is not updated once it is uploaded to the website. However, there are no clear policies on what actions the assessor takes when glaring errors are discovered.

The dataset "List of Property Sales" is continually updated between April 1st of one year and March 31st of the following year. This is because it serves as a running tally of property sales in the town. It appears to be updated every several weeks.

If the dataset becomes obsolete how will this be communicated?

By definition, both datasets become obsolete after the year they cover passes. However, this speaks to one of the main purposes of both datasets which is to facilitate a yearly assessment of real estate and tax within the town of Harpswell. Although each dataset becomes obsolete, two current datasets are released immediately following obsolescence. Thus, as a whole, the data available is continually up-to-date even if each individual dataset is not. Current datasets and the previous datasets for five years are available on the town website.

Harpswell Assessor's Office Website[7]

There is no explicit erratum for either of the two datasets.

Is there a repository to link to any/all papers/systems that use this dataset?

No repositories link to any usages of either dataset. However, this makes sense as there are no known uses outside of internal town proceedings and general public reference.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

No, there is no official mechanism to augment either dataset.

Additional Comments

The datasets appear to be exceptionally well-maintained with an update timestamp clearly visible.

Legal & Ethical Considerations

If the dataset relates to people or was generated by people, were they informed about the data collection?

Yes, when individuals submit their property tax, real estate, and real estate transactional information to the town, they are aware that it will be publicly available in the form of both datasets.

If it relates to other ethically protected subjects, have appropriate obligations been met?

N/A: Both datasets are related to humans only.

If it relates to people, were there any ethical review applications/reviews/approvals?

Unknown for either dataset but likely, as the release of these types of datasets is commonplace, and even required, across the state.

[6] Harpswell, *Contact Info*.

[7] Harpswell, *Welcome to the Harpswell Assessor's Office*

If it relates to people, were they told what the dataset would be used for a did they consent? Which community norms exist for data collected from human communications?

Yes, (See Question "If the dataset relates to people or was generated by people, were they informed about the data collection?") in the case of both datasets. Community norms for human communications are not applicable for either dataset.

If it relates to people, could this dataset expose people to harm or legal action?

There is minimal risk for legal harm as both datasets are very much public and meant to be easily accessible. However, in terms of social harm, the dataset reveals home values and the names of those who own them which could be stressful for some individuals.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

Both datasets, by their nature, only represent those who own homes. Thus, any groups underrepresented with respect to home ownership and the entire group of renters in general will be disadvantaged in terms of representation. However, in the case of this study, the group of interest *is* homeowners and thus this is not a particularly problematic issue.

If it relates to people, were they provided with privacy guarantees?

In some sense, yes because residents of the town are made aware of the public release of information policies when conducting business with real estate. However, names and addresses are not removed from associated real estate data and so that information is also publicly available, even if the owner is not a resident.

Does the dataset comply with the EU General Data Protection Regulation (GDPR) or other comparable standards?

The two datasets do not comply with the GDPR because they contain extensive personally identifying information.

Does the dataset contain information that might be considered sensitive or confidential?

The datasets both contain information that could be considered sensitive because they contain a list of names and addresses. However, the information is already publicly available.

Does the dataset contain information that might be considered inappropriate or offensive?

No, the dataset does not contain any information that could be considered offensive. However, some may consider it a faux pas to know about the home value of others in the community.

Additional Comments

Although the data is publicly available, the fact that names and addresses are included is still concerning. The town is extremely small, and the information is really only known to those involved directly in town affairs. However, a published paper would greatly increase the exposure of the dataset, potentially to those with nefarious intentions and certainly to people who have nothing to do with the town. Thus, it is worth considering redacting the names when releasing the dataset with a published paper.

Contact Information

Contact principal researcher, Shea Nelson, at shea.nelson@tufts.edu for further information and any questions, concerns, or comments.